DOCUME'T RESUME

ED 117 172 TM 005 028

AUTHOR	Abrams, Marshall D.; Cotton, Ira W.
TITLE	The Service Concept Applied to Computer Networks. Technical Note 880.
INSTITUTION	National Bureau of Standards (DOC), Washington, D.C. Inst. for Computer Sciences and Technology.
REPORT NO	NBS-TN-880
PUB DATE	Aug 75
NOTE	42p.
AVAILABLE FROM	Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (Catalog No. C13.46:880, $0.85)
EDRS PRICE	MF-$0.76 HC-$1.95 Plus Postage
DESCRIPTORS	*Computers; Delivery Systems; Display Systems; Efficiency; Information Networks; *Information Systems; Input Output Devices; Man Machine Systems; Measurement; Models; *On Line Systems; *Performance Criteria; Time Sharing
IDENTIFIERS	*Network Measurement System

ABSTRACT

The Network Measurement System (NMS) represents the implementation of a new approach to the performance measurement and evaluation of computer network systems and services. By focusing on the service delivered to network customers at their terminals, rather than on the internal mechanics of network operation, measurements can be obtained which are directly relevant to user needs and management concerns. Furthermore, the type of measurement necessary to implement this approach can be made directly, without perturbing the network system under test. This technical note introduces the service concept and other background information necessary to understand the need for and use of the NMS. The fundamental distinction between service and internal efficiency is clarified, both in general and in the environment of computer networks. A number of different measures of service are then discussed, followed by the presentation of several models of interactive use of networks. Then, the practical aspects of gathering data and applying this information to service measurement are reviewed, leading to a presentation of the NMS as it is implemented. The note ends with a discussion of applications for the NMS. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

IR

NBS TECHNICAL NOTE 880

ED117172

IM005 028

U.S. DEPARTMENT OF COMMERCE / National Bureau of Standards

The Service Concept Applied to Computer Networks

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards [1] was established by an act of Congress March 3, 1901. The Bureau's overall goal is to strengthen and advance the Nation's science and technology and facilitate their effective application for public benefit. To this end, the Bureau conducts research and provides: (1) a basis for the Nation's physical measurement system, (2) scientific and technological services for industry and government, (3) a technical basis for equity in trade, and (4) technical services to promote public safety. The Bureau consists of the Institute for Basic Standards, the Institute for Materials Research, the Institute for Applied Technology, the Institute for Computer Sciences and Technology, and the Office for Information Programs.

THE INSTITUTE FOR BASIC STANDARDS provides the central basis within the United States of a complete and consistent system of physical measurement; coordinates that system with measurement systems of other nations; and furnishes essential services leading to accurate and uniform physical measurements throughout the Nation's scientific community, industry, and commerce. The Institute consists of a Center for Radiation Research, an Office of Measurement Services and the following divisions:

Applied Mathematics — Electricity — Mechanics — Heat — Optical Physics — Nuclear Sciences [2] — Applied Radiation [2] — Quantum Electronics [3] — Electromagnetics [3] — Time and Frequency [3] — Laboratory Astrophysics [3] — Cryogenics [3].

THE INSTITUTE FOR MATERIALS RESEARCH conducts materials research leading to improved methods of measurement, standards, and data on the properties of well-characterized materials needed by industry, commerce, educational institutions, and Government; provides advisory and research services to other Government agencies; and develops, produces, and distributes standard reference materials. The Institute consists of the Office of Standard Reference Materials and the following divisions:

Analytical Chemistry — Polymers — Metallurgy — Inorganic Materials — Reactor Radiation — Physical Chemistry.

THE INSTITUTE FOR APPLIED TECHNOLOGY provides technical services to promote the use of available technology and to facilitate technological innovation in industry and Government; cooperates with public and private organizations leading to the development of technological standards (including mandatory safety standards), codes and methods of test; and provides technical advice and services to Government agencies upon request. The Institute consists of a Center for Building Technology and the following divisions and offices:

Engineering and Product Standards — Weights and Measures — Invention and Innovation — Product Evaluation Technology — Electronic Technology — Technical Analysis — Measurement Engineering — Structures, Materials, and Life Safety [4] — Building Environment [4] — Technical Evaluation and Application [4] — Fire Technology.

THE INSTITUTE FOR COMPUTER SCIENCES AND TECHNOLOGY conducts research and provides technical services designed to aid Government agencies in improving cost effectiveness in the conduct of their programs through the selection, acquisition, and effective utilization of automatic data processing equipment; and serves as the principal focus within the executive branch for the development of Federal standards for automatic data processing equipment, techniques, and computer languages. The Institute consists of the following divisions:

Computer Services — Systems and Software — Computer Systems Engineering — Information Technology.

THE OFFICE FOR INFORMATION PROGRAMS promotes optimum dissemination and accessibility of scientific information generated within NBS and other agencies of the Federal Government; promotes the development of the National Standard Reference Data System and a system of information analysis centers dealing with the broader aspects of the National Measurement System; provides appropriate services to ensure that the NBS staff has optimum accessibility to the scientific information of the world. The Office consists of the following organizational units:

Office of Standard Reference Data — Office of Information Activities — Office of Technical Publications — Library — Office of International Relations.

---

[1] Headquarters and Laboratories at Gaithersburg, Maryland, unless otherwise noted; mailing address Washington, D.C. 20234.
[2] Part of the Center for Radiation Research.
[3] Located at Boulder, Colorado 80302.
[4] Part of the Center for Building Technology.

The Service Concept Applied to Computer Networks

Marshall D. Abrams and Ira W. Cotton

Institute for Computer Sciences and Technology
National Bureau of Standards
Department of Commerce
Washington, D.C. 20234

U.S. DEPARTMENT OF COMMERCE, Rogers C. B. Morton, *Secretary*
NATIONAL BUREAU OF STANDARDS, Ernest Ambler, *Acting Director*

Issued August 1975

Library of Congress Catalog Card Number: 75-600056

National Bureau of Standards Technical Note 880

Nat. Bur. Stand. (U.S.), Tech. Note 880, 38 pages (Aug. 1975)

CODEN: NBTNAE

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON: 1975

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402
(Order by SD Catalog No. C13.46:880). Price 85 cents. (Add 25 percent additional for other than U.S. mailing).

CONTENTS

	Page
INTRODUCTION	1
SERVICE VERSUS EFFICIENCY	3
MEASURES OF SERVICE	5
Response Time	5
Psychological Considerations	6
Other Factors	7
Throughput	7
Cost-Effectiveness	8
Cost-Benefit Considerations	9
CONCEPTUAL MODELS	10
The Stimulus-Response Model	11
Communications Conventions	13
Interleaving Conventions	13
The Stimulus-Acknowledgement-Response Model	15
A Full-Duplex Model	16
MEASUREMENT OF SERVICE	16
What to Measure	18
Where to Measure	20
How to Measure	20
Network Measurement Machine	21
Data Analysis	23
Subsets	28
USES OF THE NETWORK MEASUREMENT SYSTEM	28
Procurement	29
Operation	29
Applications	30
CONCLUSIONS	32
FOOTNOTES	32
BIBLIOGRAPHY	33

The Service Concept Applied to Computer Networks

Marshall D. Abrams and Ira W. Cotton

Abstract

The Network Measurement System (NMS) represents the implementation of a new approach to the performance measurement and evaluation of computer network systems and services. By focusing on the service delivered to network customers at their terminals, rather than on the internal mechanics of network operation, measurements can be obtained which are directly relevant to user needs and management concerns. Furthermore, the type of measurement necessary to implement this approach can be made directly, without perturbing the network system under test.

This technical note introduces the service concept and other background information necessary to understand the need for and use of the NMS. The fundamental distinction between service and internal efficiency is clarified, both in general and in the environment of computer networks. A number of different measures of service are then discussed, followed by the presentation of several models of interactive use of networks. Then, the practical aspects of gathering data and applying this information to service measurement are reviewed, leading to a presentation of the NMS as it is implemented. The note ends with a discussion of applications for the NMS.

Keywords: Computer networks; interactive computing; man-machine interaction; performance measurement; time-sharing

INTRODUCTION

In recent years, increasing numbers of people have begun to use computers through interactive terminals in a conversational mode. Rather than submitting jobs in a batch mode and waiting hours (or days) for the results, users interact with the computer continuously on a transaction-oriented basis. This trend gives every evidence of

continuing, and it appears that this will be the predominant form of computer access in the future. This type of computer access is characteristic of most current computer networks.

As discussed here, a computer network consists of one or more computers, one or more human users employing an interactive terminal, and the communications facilities which interconnect them [BL74]. The user of a computer network is a customer for computer services. The communications portion of the network is only the delivery mechanism for the service which the user observes and evaluates. However, as a delivery mechanism, the communications facility may have a significant effect on the service received by the user at his terminal.

Looking at the population of computer terminal users today and projecting into the future, we see that most of them are not programmers or computer specialists; rather, they are stockbrokers, reservation clerks, librarians, ... -- in short, they are applications-oriented, rather than computer-oriented. Most of these applications are themselves service-oriented, so it is not surprising that the terminal users are only interested in the computer in terms of the service it can provide to them.

The user is typically unconcerned with the technical details of how these services are produced or how they are delivered. The user probably becomes concerned with these factors only when they somehow fail to meet his requirements. It is particularly important to note that the customer for computer services is unconcerned with differentiating between the computer system and the communications system. When a computer network is used, both contribute to the service which is employed. The services rendered by a computer network are of value to a user only at the point at which the service is delivered. For conversational computer utilization, service is delivered at and through the terminal.

When such customers solicit computer services, they are neither interested in the internal operation of the central computer facility nor in the internal performance of the communication facility; they are interested only in the level of computer service delivered at the terminal. To intelligently compare alternative suppliers, they need quantitative measures of this service. However, the mainstream of interest in computer system performance evaluation in past years has focused on the evaluation of internal computer system operation, rather than on this problem.

In this technical note, we introduce an approach to the

evaluation of computer network performance which does focus on the service provided to the terminal user. We explain how this approach differs from previous approaches, we introduce criteria for and measures of computer service, and we describe a new tool which has been designed and implemented to acquire network usage data and interpret it in terms of these measures. This tool, the Network Measurement Machine (NMM), is part of a larger Network Measurement System (NMS) which includes a library of data analysis programs and which can be used for a variety of measurement applications. We conclude with a discussion of the present and planned applications of the Network Measurement System.

SERVICE VERSUS EFFICIENCY

Much attention has recently been directed at measurement and related quantitative evaluation techniques as an important part of the selection and improvement of the hardware and software of computer systems. (An extensive bibliography may be found in Miller [MIe72]). The term "performance evaluation" is frequently used to refer to such investigations. Performance evaluation has primarily been concerned with the internal operating efficiency of computer systems, measured in terms of throughput, or the quantity of work which can be processed in a given time interval. Factors such as Central Processor Unit (CPU) utilization and peripheral device utilization have been measured by hardware monitors. Similarly, components of operating system and user software have been measured by software monitors. The raw data from these monitors is processed into performance reports. Analysis of these reports has provided insight into hardware/software modifications designed to improve the cost-effectiveness of the computer installation. The usual types of positive action taken as a result of such internal performance evaluation studies are a reconfiguration of the hardware, a change in system tables, or an improvement to a software module. Typical major changes might include the procurement of an additional disk drive or replacement of the scheduling algorithm. The overall results of such changes can be quite dramatic increases in system throughput.

The basis for measuring success in system changes of this type is some measure of the cost-effectiveness of the system before and after the changes. The cost-effectiveness calculations are usually based on the time to run some given set of jobs (a "benchmark") and the cost (purchase or monthly rental) of the particular configuration. Thus, cost-effectiveness is really a ratio of throughput to cost, and the analyst is presumed to be indifferent to a reduction

in cost for identical throughput, or increased throughput for identical cost. The time to process the benchmark is important only as a measure of throughput.

Individual users, however, may not be indifferent to this tradeoff of cost versus time. For them, service is the key factor, not throughput. In contrast to the evaluation of efficiency, which is concerned with the time and cost to run a group of jobs, service evaluation is concerned with the time and cost to run each individual job. Actually, the times of concern are of different types. Efficiency is concerned only with running or execution times, while service evaluation is concerned with total elapsed time. (In most multiprogramming systems, the elapsed time for a job is considerably greater than the run time). The evaluation of efficiency is thus based on the measurement of the internal functioning of a computer system as a whole, rather than of its external manifestations, taken individually.

The goals of improved efficiency and improved service may well be at odds with one another. An improvement in internal performance does not necessarily imply an improvement in service. Indeed, the opposite may actually be true. The principle of most multiprogramming systems is to process one program while others are in a wait state. This serves to improve efficiency. However, there is a delay in returning control to another program when the wait state is terminated; this delay represents a decrease in service to the waiting program. With many multiprogrammed interactive users, this delay may be quite significant. Frequently, it is only possible to improve service at the expense of efficiency.

For measuring service, the methodology of cost-benefit analysis is more appropriate than that of cost-effectiveness. Cost-benefit analysis recognizes that jobs may have a time-value, and that a simple average of throughput is an inadequate measure. In a sense, measures of service must also be concerned with the standard deviation of run or response time. When measuring performance, jobs run more quickly than average may cancel out the effect of more slowly run jobs; in measuring service there are only two types of jobs - acceptable and unacceptable. No further merit is ascribed to jobs run more rapidly than what is deemed "acceptable."

Cost-benefit analysis also recognizes that there may be qualities of service which are not directly measurable in terms of cost or time to run the job. Examples of such qualities are accuracy, comprehensibility, ease of use, dependability and the like. Though possibly not directly measurable in terms of time or cost, they may be considered

through their indirect effect on either the cost or benefit side of the analysis. For example, an inaccurate system may increase costs by requiring transactions (or entire jobs) to be rerun when errors occur. Similarly, an easy to use system is more beneficial to use than one that is not so, because of the costs incurred as a result of user errors or delay.

MEASURES OF SERVICE

In this section we review the basic units of measurement for interactive computer service. Response time is a fundamental unit, but it has several interpretations, and the implications of a particular value are not clear. Throughput is another similar unit of measurement, though most suitable for batch applications.

Based on measures of response time and/or throughput, network configurations can be manipulated in a search for the most cost-effective. However, cost-effectiveness analysis tells nothing about the value of any improved performance. For a truly global figure of merit, cost-benefit analysis is the proper approach.

Response Time

The most common measure applied to conversational systems is the "response time." This time is most frequently defined to be the elapsed time from the end of the stimulus to the beginning of the response. Although intuitively appealing, this measure of response time can lead to quite erroneous conclusions when blindly applied. To ameliorate these (as yet unnamed) difficulties an alternate definition of response time is the elapsed time from the final printing character in the stimulus to the first printing character in the response. Other measures of the system's responsiveness have been proposed [AB73].

The general goal of most system designers has been to achieve the best (shortest) response time possible. Too often, systems have been designed without a particular response time goal in mind, possibly because it was not known what response time was needed. This determination is most properly the domain of engineering-oriented psychologists, some of whom have begun to address the problem in recent years.

Psychological Considerations

In 1968 two papers appeared which addressed the psychology of user-computer conversational interactions. Neither of these papers reported original experimental research, but rather applied previous work to the (then) relatively new phenomenon of conversational computing. Concepts such as sets of response times, the acceptability of a given response time relative to the task to be performed, the tendency of a user to reduce his concentration if the response time is too long, and the undesirability of unpredictable response times are discussed. Carbonell et. al. [CA68] went on to develop a mathematical model of user acceptability. Miller [MIr68] analyzed various sample types of conversational computing and presented estimates of acceptable response times for a variety of circumstances based on consideration of operating needs, psychological needs, expectancies, activity clumping, psychological closure, human short-term memory and psychological step-down discontinuities with increasing response time, (i.e. loss of attentiveness caused by increasing response time). This paper appears to be the source of the often quoted (partially out of context) statement: "For good communication with humans, response delays of more than two seconds should follow only a condition of task closure as perceived by the human, or as structed for the human." Another psychological measure, the rate at which users learn to use a system, was proposed by Jutila and Baram [JU71] in 1971. They presented actual empirical results in a mathematical model formulation.

The preceding three papers cited may represent good psychology. The application of these and similar theoretical predictions and tentative experimental observations may or may not represent good management. The extrapolation that keeping the individual user unfrustrated is good for the organization appears to be a variant on the theme that "contented cows give sweet milk." Reasoning by analogy is always risky and should be subject to independent confirmation whenever possible. One experimental study at the RAND Corporation [BOE71] cast doubt on this hypothesis. In the experiment, two groups were given identical problems to solve. In one group the computer responded as rapidly as possible, while in the other group the response time was constrained to a certain minimum. That is, at times the computer responded more slowly than it could have, in an effort to force users of that group to reflect on what they were doing. Better or faster solutions to the problem were reported for the latter group, though accompanied by widespread feelings of frustration.

What this implies for management is not clear. In terms of overall performance, it has been demonstrated that faster is not always better. On the other hand, in times when qualified personnel are in short supply it may not pay to frustrate them unnecessarily; they may leave. Clearly, additional work is necessary before the optimal system response time is known.

Other Factors

A number of other factors besides responsiveness are relevant to the service provided by a network to interactive users. Grubb and Cotton [GR75], in surveying the general parameters of data communications facilities used by information processing systems, suggested nine parameters for comparing systems:

1. Transfer Rate
2. Availability
3. Reliability
4. Accuracy
5. Channel Establishment Time
6. Propagation Delay
7. Turnaround Time
8. Transparency
9. Security

Two of these factors, transfer rate and propagation delay, directly influence response time. Channel establishment time could be measured with the tools to be described here. Turnaround time is only applicable to half-duplex communications and so not of general concern. The other factors -- availability, reliability, accuracy, transparency and security -- do affect the service which is delivered to network users. These factors are not addressed by the measurement approach described in this technical note.

Throughput

In conventional batch operation the measure most closely related to response time is "turn-around" time, which is the elapsed time from the submission of a card deck (or its equivalent) until printout is obtained. This is a very important parameter to the individual user, since it regulates his access to the computer. Another measure applied to batch operation is the "throughput," which is the ratio of work per unit time. In a mono-programming environment when the unit of work is taken to be the card deck (or equivalent), throughput is the reciprocal of

turn-around time. In a multi-programming environment, throughput may also be expressed in jobs per hour (or similar units), but the overall system performance is not directly calculatable from that of the individual jobs. Furthermore, as identified by Kamerman [KA69], in an early paper setting the stage for the evaluation of conversational systems by IBM, the workload in each job is a variable that must be specified in order to calculate throughput. Assuming that this is done, Kamerman states that "throughput provides the best measure of conversational system performance." Applied to the iterative design of a single conversational system, these remarks serve to introduce concepts of cost-effectiveness.

Cost-Effectiveness

Cost-effectiveness analysis is a technique used to compare alternative systems where the benefits to be derived cannot be determined quantitatively. The object of cost-effectiveness analysis is to select either the system with lowest cost for a given level of performance or greatest performance for a given level of cost. The method requires that either the costs or the performance associated with all the systems under consideration be equal. If both costs and performance are different for any two systems to be analyzed, then cost-effectiveness analysis is impossible since the parameters have different dimensions and may not be compared.

As defined, the application of cost-effectiveness analysis is fairly straightforward. First it must be decided which parameter to hold constant -- cost or performance. Cost is frequently held constant when there is no absolute level of performance which is required. In this case, a fixed amount of money is made available and the analysis seeks the system with which offers the greatest performance. A measurement criteria is selected and the systems are evaluated according to that measure. The system which measures the greatest --for identical cost-- is the most cost-effective. Tradeoffs between system components may be investigated by varying the mix of components subject to the constraint that total system cost remain the same.

The alternative approach is to specify in advance the performance required of the system. The system is selected which offers at least that level of performance for the least cost. Performance in excess of the specified capabilities is ignored in the analysis. For such systems, tradeoffs focus on the possible reductions of performance (and with it, cost) to the point where performance just equals that required.

Cost-Benefit Considerations

Cost-effectiveness analysis is valuable because it shows how to best allocate limited resources so as to optimize some specific performance objectives. However, such an analysis is inadequate for making an investment decision. Such decisions require that total system cost be compared with the value of the benefits provided by the system. This requires that acceptable levels of performance be defined, so that a point on the cost-performance curve (corresponding to a particular system) can be selected. Comparing total costs with the benefits to be derived from the system determines whether or not the system selected is cost-beneficial. Unfortunately, the problem is rarely expressed this way in the literature [CO74].

The real problem with cost-benefit analysis arises on the benefit, not the cost, side. It is important to include all relevant costs, both direct and indirect, but this is a relatively straightforward accounting problem. More difficult is determining the benefits from some given level of system performance. For our case, it is necessary to assign some value to the results of computation in order to perform a cost-benefit analysis of computing services. One such analysis has been performed by Streeter [ST72] for batch and conversational systems supporting scientific applications.

Streeter recognizes that the direct invoice costs are not the only costs associated with computer utilization. There are many other costs borne by the user organization, including the salary of the problem solver and supporting personnel such as programmers and key-punch operators, the cost of communications, the cost of supplies, and the cost of equipment (rental or ammortization). Streeter further suggests that computational results have a time value which is highest when the computation is requested and ever decreasing thereafter. Estimating this value and the time-dependent function is admittedly complex. The decrease in the value of the output may be treated as an additional cost to be added to the costs already itemized. Including this cost factor is one way of directly linking response time and turnaround to cost calculations.

Streeter also recognizes that evaluation of the time value requires a "subjective judgement of diminution of the relative value of a job to the user as a function of system response time." The concept of value to the user represents a departure from usual system efficiency measures such as system throughput. For transaction systems such as

point-of-sale and airline reservations, assignment of a time-value was (or should have been) part of the original system design, and is, therefore, a known value.

In many circumstances, however, the time-value of a computation will not have been previously determined. Not many managers have sufficient experience to be able to assign a realistic estimate. In these cases the time value term may have to be neglected. It should be recognized that omitting the time-value term may well change the results of an evaluation. When a comparison is made among alternate suppliers of computer services, this omission could lead to a sub-optimum selection.

Considering all the factors which have been considered, we have, for the total cost evaluation:

(total problem solving cost) = (computer systems cost)
+ (communications cost) + (cost of users time)
+ (cost of delay) + (auxiliary charges)

This formulation expresses the relationship between the various cost factors and the parameters of network operation. Not addressed is the question of the original value of the computation itself to the organization; this presumably was addressed in some fashion when the basic decision to computerize was made. For organizations contemplating replacing manual operations with network services, this calculation might be of paramount concern; here, unfortunately, Streeter's work offers little assistance.

CONCEPTUAL MODELS

Complex phenomena are often best understood in terms of models. Model making and testing is a fundamental part of the "scientific method" of the natural sciences, with which computer science is often grouped. For the purposes of this discussion, it is sufficient to note two points about conceptual models: (1) The simplest model that incorporates all the observations describing a phenomenon is employed. (2) When interpretation of data according to the model produces results at variance with physical reality, it is time to change the model.

In order to investigate the service delivered by interactive systems, a set of computer programs have been written which constitute the embodiment of several models. These programs accept as input the measured data and calculate various items of derived data required by that model. Statistical analysis of this measured and derived

data is also performed, as is report generation. An understanding of these models is fundamental to an appreciation of our approach to network measurement and evaluation.

The Stimulus-Response Model

The simplest model is the stimulus-response model, which views the conversation between user and system as a sequence of individual transactions which are clearly delimited. In this model the human is considered to be in control; his inputs to the computer are termed "stimuli." The computer is considered to be the object of study (and use); its outputs to the human are termed "responses." The simplest pattern of computer usage is a transaction in which the user issues a stimulus and waits until the computer issues a response. It may be assumed that the next stimulus is dependent upon the contents of the preceding response. This transaction, or stimulus-response pair, is known as a "message group." Figure 1 illustrates several message groups from an actual conversation.

Because the stimulus-response model is based on the analysis of message groups, special "boundary condition" problems may arise with the first and last communication with the computer. These message groups may not be complete. The first message group may be incomplete when the computer produces output first, without waiting for data communication from the user. This is a not infrequently encountered design on the part of the initialization and login procedures. In this case we consider that the user has indeed issued a stimulus by virtue of employing the communications aspect of the network to establish communication with the computer. The last message group may likewise be incomplete. The computer may not respond to the stimulus which constitutes the termination of the data communication.

Fortunately, for our purposes these boundary conditions do not constitute the major interest nor the bulk of the analysis. They can be simply explained and accounted for in the analysis. For studies concerned with communications line utilization, response time and throughput, these boundary conditions may be deleted winout loss of generality. Other studies concerned with communications initialization may wish to concentrate attention on these boundary conditions instead of the remainder of the conversation.

```
CONVERSATION RECORD OF FILE 4196#1-1108.HDX;1

<S=STIMULUS, R=RESPONSE, A=ACKNOWLEDGEMENT, E=ECHO>
<SD=STM DLY TIME, ST=STM XMIT TIME, SC=STM CHAR COUNT, SR=STM TRANS RATE>
<AD=ACK DLY TIME, AT=ACK XMIT TIME, AC=ACK CHAR COUNT, AR=ACK TRANS RATE>
<RD=RESP DLY TIME, RT=RESP XMIT TIME, RC=RESP CHAR COUNT, RR=RESP TRANS RATE>
<SRD=STM-RESP DLY, SI=STM INTER ARRIVAL TIME>

S 1     XXXXX
A 1     <CR><LF>*2>
R 1     UNIVAC 1108 TIME/SHARING EXEC <SP>*2>VERS 225 UPDATE B<CR>
        <LF>*2>
                        ST:     22.0    SC:     63      SR:     2.9
        AD:     0.2     AT:     0.2     AC:     3       AR:     17.5
        RD:     .001    RT:     1.8     RC:     51      RR:     28.4
        SRD:    0.4     SI:     0.0


RECORDING TIME: MONDAY, JULY 15, 1974 9:02AM-EDT


S 2     @RUN AAAAA,BBB-CCC,2<CR>
A 2     <LF><CR>
R 2     DATE: 071574 <SP>*6>TIME: 090315<CR><LF><DEL>
        SD:     4.4     ST:     13.3    SC:     31      SR:     2.3
        AD:     .036    AT:     5.9     AC:     2       AR:     0.3
        RD:     0.2     RT:     1.1     RC:     33      RR:     28.9
        SRD:    6.1     SI:     28.6

S 3     @ED,U ELEMENT.ELT1<CR>
A 3     <LF><CR>
R 3     ED 13.00-07/15-09:03-(27,28) <SP>*2><CR><LF>EDIT <SP>*2><CR>
        <LF>0:
        SD:     0.7     ST:     4.6     SC:     19      SR:     4.1
        AD:     .036    AT:     3.0     AC:     2       AR:     0.7
        RD:     0.2     RT:     1.9     RC:     42      RR:     22.5
        SRD:    3.3     SI:     21.2

S 4     L OUT2<CR>
A 4     <LF><CR>
R 4      <SP>*9>J <SP>*7>OUT2 <SP>*3><CR><LF>101:
        SD:     1.3     ST:     2.2     SC:     7       SR:     3.1
        AD:     .036    AT:     0.5     AC:     2       AR:     3.7
        RD:     0.2     RT:     1.2     RC:     30      RR:     24.1
        SRD:    0.7     SI:     11.0
```

Figure 1 Partial Conversation Record

Communications Conventions

Within the body of the conversation the identification of responses becomes intertwined with the communications convention for the relationship between the keyboard character struck by the user and the printing on the terminal printer./1/ When communications employs the echo-plex convention, every character transmitted by the user may be considered a stimulus and every echo thereof from the network, a response. Again depending upon the purpose of the study, this may or may not be a valuable model representation. For a study concerned with line utilization or echo propagation characteristics this would be the valid and useful manner in which to consider the conversation. For a study concerned with a more macroscopic view of the conversation, the echoed characters represent an unwelcome detail.

To cope with this problem, we have provided alternatives in the analysis programs which implement the model. Optionally, the echoed characters may be discarded, thereby transforming an echo-plex conversation into a half-duplex one. When this transformation is applied to a full-duplex conversation, those characters which are echoed exactly as transmitted are correctly deleted. Characters which are transformed for echoing, the usual case when a non-printing control character is echoed, are not deleted. The original transmitted character then becomes the last character in the stimulus and the echo constitutes the beginning of the response.

Interleaving Conventions

In addition to communications conventions, computer hardware and operating systems also enforce conventions as to stimulus interleaving. At one extreme, interleaving is prevented by locking the keyboard until the response is complete. Equally severe is the disregarding of additional stimuli received. More flexible systems accept multiple stimuli by queuing them. Such queuing is independent of the communications conventions discussed above. When modeled by the stimulus-response model, queued stimuli appear as a single stimulus. If all the responses to queued stimuli are contiguous in output they appear as a single response. Even more complicated is the situation where the user stimuli are intermixed with responses to previous stimuli (i.e. a response does not immediately follow its stimulus). Figure 2 illustrates this problem. Semantic content analysis is required to match stimulus and response in this situation.

a) Normal Sequence

```
User:      5+7<CR>      Message Group 1
Computer:  12                   "
User:      1+2<CR>      Message Group 2
Computer:  3                    "
User:      9+9<CR>      Message Group 3
Computer:  18                   "
```

b) Queued Stimuli

```
User:      5+7<CR>      Message Group 1
User:      1+2<CR>             "
User:      9+9<CR>             "
Computer:  12                  "
Computer:  3                   "
Computer:  18                  "
```

c) Interleaved Stimuli and Responses

```
User:      5+7<CR>      Message Group 1
User:      1+2<CR>             "
Computer:  12                  "
User:      9+9<CR>      Message Group 2
Computer:  3                   "
Computer:  18                  "
```

Figure 2. Full-Duplex Possibilities

Our analysis programs currently do not provide this capability.

The Stimulus-Acknowledgement-Response Model

Interpretation of data from certain systems according to the stimulus-response model led to the anomaly that computer conversations which left the human user with a subjective evaluation of very slow response appeared to have very fast response when analyzed. Study of the data revealed that we had not completely considered communications conventions. ASCII/2/ specifies the format effector interpretation of the carriage return (CR), but not its conventional use in conversational applications. Many computers employ the convention that a line constitutes a unit of input and that a line is terminated by a CR. Since the CR (only) returns the print position to the first column of the current line, the computer issues a line feed (LF) following receipt of a CR to move the print position to the following line. (Some computers require that the sequence CR LF terminate the line, thus obviating the problem). Other control characters may accompany the LF, but their presence generally is not even recognized by the user since they cause no printing or motion of the print mechanism.

According to the stimulus-response model, these non-printing characters constitute the beginning of the response; by subjective human judgment they do not. Their presence is important, however, in that they provide feedback which reassures the user that the computer is still functioning. Whenever the user types a CR he is reassured that the network computer is still serving him by the action of the LF. Nevertheless, this LF does not constitute a meaningful response to the stimulus. In psychological terms, the LF provides partial closure, but the user is still waiting for complete closure to be provided by the response.

To account for this anomaly, the stimulus-response model was modified. A new state was introduced called the "acknowledgement;" this model is accordingly called the "Stimulus-Acknowledgment-Response" (SAR) model. Operationally, we first defined the acknowledgement as being composed of all the initial non-printing characters which are output by the computer following a stimulus. As in the case of the LF, the ackowledgement provides a partial closure by reassuring the user of the network's continuing ability to serve him.

Since the concept of an acknowledgement is based on psychological considerations, its content must be defined

subjectively. Once defined, the acknowledgement can be recognized by an algorithm. We have two extensions to the definition of the acknowledgement. First, we view the ackowledgement as being time-dependent. we may therefore consider the acknowledgement to be terminated by the lapse of some specified time period during which there is no transmission from the computer. Second, we may define the response as beginning with the first meaningful character printed by the computer. In particular, a fixed heading on all output may be considered as part of the acknowledgement rather than as part of the response.

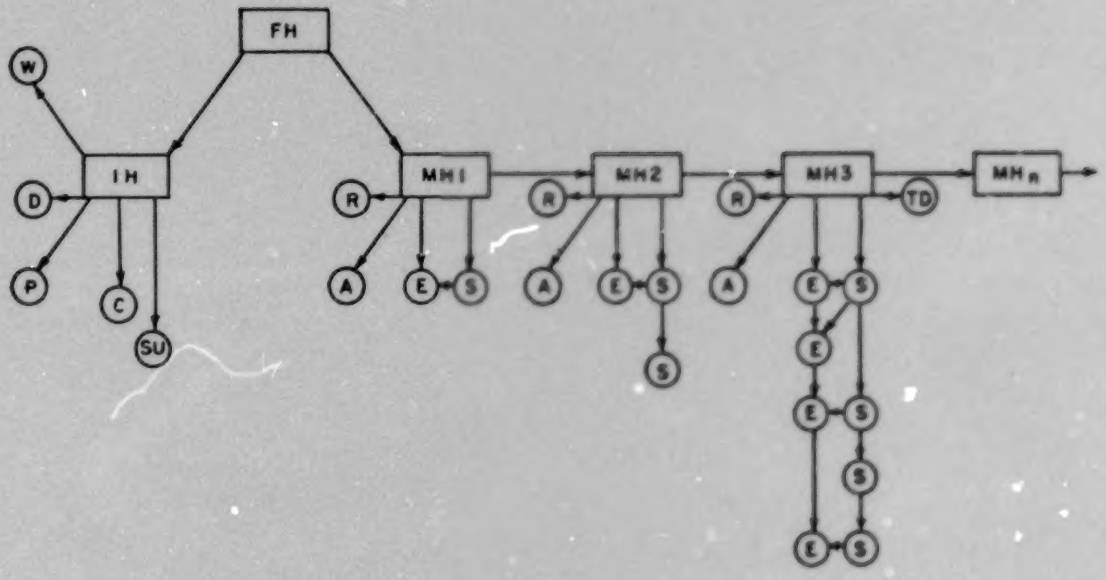
Some networks from which we have collected data appear to construct an entire output buffer before initiating any output to the user. The buffer contents are then all transmitted to the user at the maximum network transmission rate. In such a case, even though the formal definition of an acknowledgement is fulfilled, the psychological function is not. We have therefore implemented an option in our programs which suppresses the acknowledgement state. It should also be noted that in some cases the only output from the computer consists of the LF and other nonprinting characters. When this situation arises, we consider the output to be response rather than acknowledgement.

A Full-Duplex Model

The half-duplex SAR model described above is easily extended to incorporate full-duplex conversations. If we conceptualize the stimulus-acknowledgement-response model as being linear, the full-duplex model is represented by a directed graph as shown in Figure 3. For each character of the stimulus there is an associated echo. Furthermore, there may be sub-stimuli to which there will be sub-acknowledgements and sub-responses (e.g. multipart commands). The complexity of this model makes it possible to analyze data according to various models, since it completely retains all the measured data along with any analyst-introduced relationships among the data.

MEASUREMENT OF SERVICE

Thus far we have discussed the concept of service as applied to a computer network, measures of such service, and conceptual models which help us to understand the man-computer interaction. Now we turn our attention to the measurements themselves.



IH	= Information Header
FH	= File header
S	= Stimulus
E	= Echo
A	= Acknowledgement
R	= Response
TD	= Time-date indicator
MH	= Message group header
W	= Who did the conversation
D	= Date of the conversation
P	= Place of conversation
C	= Computer(s) used in the conversation
SU	= Subset Description

Figure 3. Directed Graph of Full Duplex Data Index File Structure

What to Measure

The identification and selection of the data to be collected entails a matching of what can be collected with what needs to be collected (for a given purpose). What can be collected is generally dictated by the nature of the physical system under investigation and by the capabilities and limitations of existing or planned measurement instruments. What needs to be collected is determined from the model which has been developed and from the capabilities of existing analysis procedures. In short, the data to be collected are the intersection of what can be collected and what should be collected.

In the data communications area, the basic events to be observed are the occurrence of bits on a channel in some time sequence. With proper synchronization of the measurement instrument (requiring knowledge of the transmission rate on the channel), appropriate clusters of bits may be recognized and recorded as characters. Along with each character can be recorded its time of occurrence (according to some external clock), and, if the channel under investigation is bidirectional, the source of the character.

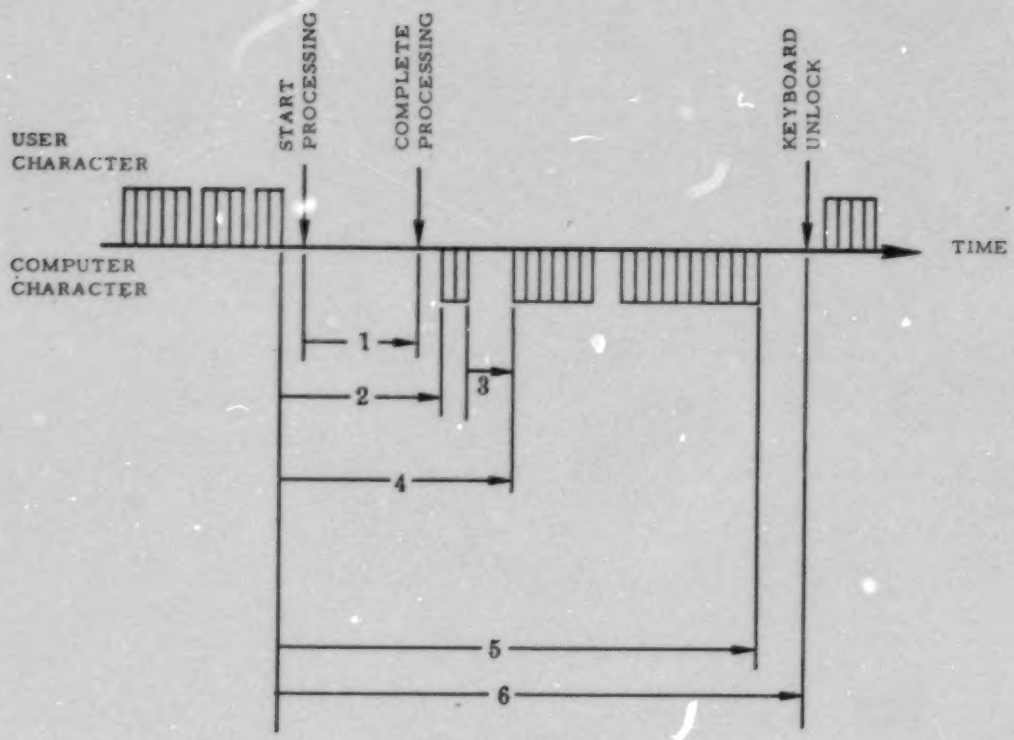
For the stimulus-acknowledgement-response model which has been described, nine relevant parameters may be derived from the basic recorded data:

stimulus delay time
stimulus transmit time
stimulus character count
acknowledgement delay time
acknowledgement transmit time
acknowledgement character count
response delay time
response transmit time
response character count

We are presently considering expanding the treatment of response delay times. The existence of various proposed measures of system responsiveness [CA68,MI68,AB73,KA69,BO74] also suggests that we ought to present more than one such measure. The following intervals are under consideration:

(1) end of stimulus to beginning of response;
(2) end of stimulus to first meaningful character in the response;
(3) end of stimulus to end of response.

These various measures are illustrated in figure 4. Other measures are possible and can easily be derived from the basic data.



1. Inquiry Response Time [Silverman, 1973]
2. Response Time, System Responsiveness [Kamerman, 1969]
   Stimulus-Acknowledgement Delay (this paper)
3. Acknowledgement-Response Delay (this paper)
4. Stimulus-Response Delay (this paper)
5. Turnaround Time, Response Time [Boehm, 1971]
6. System Response Time [Boies, 1974]

Figure 4. VARIOUS MEASURES OF SYSTEM RESPONSE

A record of each character with its source and time of occurrence is all that is necessary to completely reconstruct the physical events being observed, namely the conversation. For some applications some of this information may be superfluous. For example, all that is needed to compute throughput statistics is a count of the number of characters observed in a given time interval. For other applications, however, additional data may have to be derived from the basic recorded data. To analyze conversations in terms of the models which have been described, it is necessary to assign different character sequences to the various states of the model, and to compute the time intervals between the boundaries of these states.

Where to Measure

Although measurements may be taken anywhere in the computer network, certain points are favored because of their convenience or their logical relationship to the measurement objectives. When the measurement objective is to determine the quantity or quality of service rendered to the user at his terminal, the best point of measurement is at the terminal itself, or at least at the terminal's interface to the network. When the objective is to determine the demands placed on a network computer and its responses to those demands, the best point of measurement is in that computer itself or at the interface between the computer and the network. When the objective is to determine the effect of the network on demands and service, both of the preceding points of measurement may need to be employed, preferably together and simultaneously.

Unfortunately, practical limitations frequently interfere with the selection of a point of measurement. The (non)portability of the measurement device exerts a strong influence on the point of measurement, as does the ease of attachment to the system being measured. For certain applications, such as measuring the delay introduced by a particular communications system, simultaneous measurement of two different data paths may be necessary. Thus, the point of measurement may often be at some intermediate point in the network. We are led to seek a general device which may be used in any of these modes, which is easy to attach, and which does not perturb the system under test.

How to Measure

When the point of measurement is at the interface between the computer and the network, measurement could be performed

within the computer itself. Certainly all the information needed is available [B074]. However, to collect this information in the form desired would require additional software (and possibly hardware) in the host computer. This approach is undesirable because it perturbs the system being measured.

When the point of measurement is elsewhere, a special measurement device must be provided. Once developed, such a device could also be used at the interface between the computer and the network. This approach has the appeal of generality as well as overcoming the previous objection to perturbing the system being measured.

We have constructed a minicomputer-based data acquisition system suited to use for any of the measurement points which have been discussed. The data are not analyzed as they are collected; rather, they are stored on magnetic tape and processed later on a large scale machine with special data analysis routines. We call the data acquisition system, with its minicomputer hardware and software and special interfaces, the "Network Measurement Machine" (NMM). The data analysis routines are referred to as such, and the entire complex for data acquisition and analysis is referred to as the "Network Measurement System."

Network Measurement Machine

The Network Measurement Machine (NMM) is the data acquisition part of the Network Measurement System (NMS). It is a minicomputer-based system which employs standard and special purpose hardware under the control of a specially written operating system.

The NMM is implemented on a Digital Equipment Corporation PDP-11/20. The hardware currently in use is shown in figure 5. The standard hardware includes a high precision Programmable clock and a set of communications interfaces. Special hardware is employed to connect the NMM to the network which is to be measured. This includes portions of a general purpose data intercommunications system as well as an automatic calling unit and line selector system for computer-controlled originating and answering of data calls via common carrier communications facilities.

The special operating system is a real-time interrupt-driven scheduler incorporating various drivers and handlers for the standard and special purpose interfaces attached. It is written in assembly language and used the manufacturer's regular software in its preparation. Programming is modular,

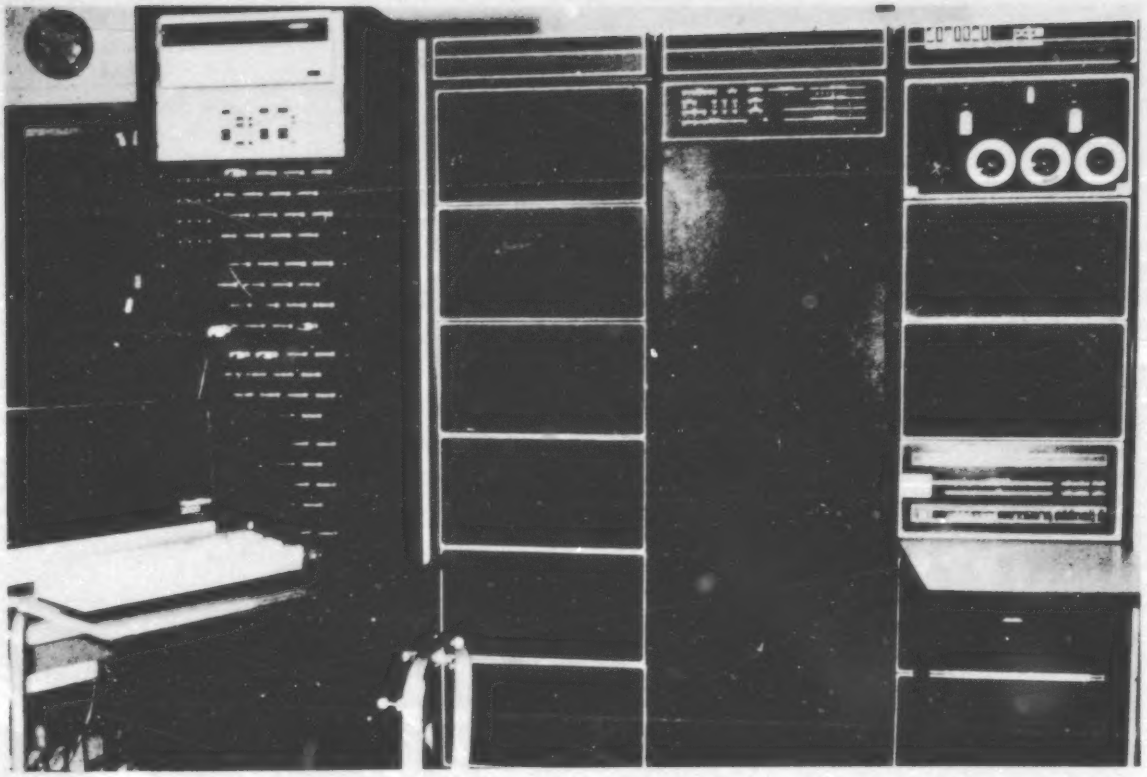


Figure 5. The Network Measurement Machine

with extensive use of semaphores for task synchronization. A command interpreter serving the operator's console makes it possible to view the status of the NMM as well as selected ongoing communication.

As described in much more detail by Rosenthal, Rippy and Wood [RO75], the function of the NMM is to receive characters from both computer and user and to record these characters along with descriptive information as to the source of the character, the time at which the character occurred, and the state of the communications line at the time the character existed. All of this information is presently recorded on 7-track magnetic tape for subsequent processing by the NMS data analysis programs, as shown in figure 6.

Data Analysis

After collecting a sample of conversations on magnetic tape (in terms of characters, time and source), it is necessary to process this "raw" data to obtain summary statistics from which conclusions can be drawn. As with many other types of measurement, the acquisition of data has proven to be simpler than its interpretation. A set of data analysis programs has been written to process the data according to one or more of the models which have been described. In order to appreciate the significance of any of the summary statistics which may be derived, it is necessary to understand the particular model being used, and the way in which the data are manipulated according to this model. We present a brief outline of the data analysis procedures; complete documentation may be found elsewhere [WA75].

Since the data acquisition process in the Network Measurement Machine may have involved the intermixing of multiple conversations, the first step must be the demultiplexing of these conversations. This requires the recognition of conversation boundaries. For conversations which are recorded through different interfaces there is no problem; each data stream is labeled as it is recorded, so that the various streams can easily be unraveled. The real problem arises with sequential conversations recorded through the same interface.

Depending on the operating procedures of the users being measured and the characteristics of the hardware used to acquire data at the various measurement points, there may or may not be a physical indication when a conversation terminates. That is, not all users hang up the line between different conversations, and not all measurement hardware is equipped with carrier detect. In such cases, it is necessary

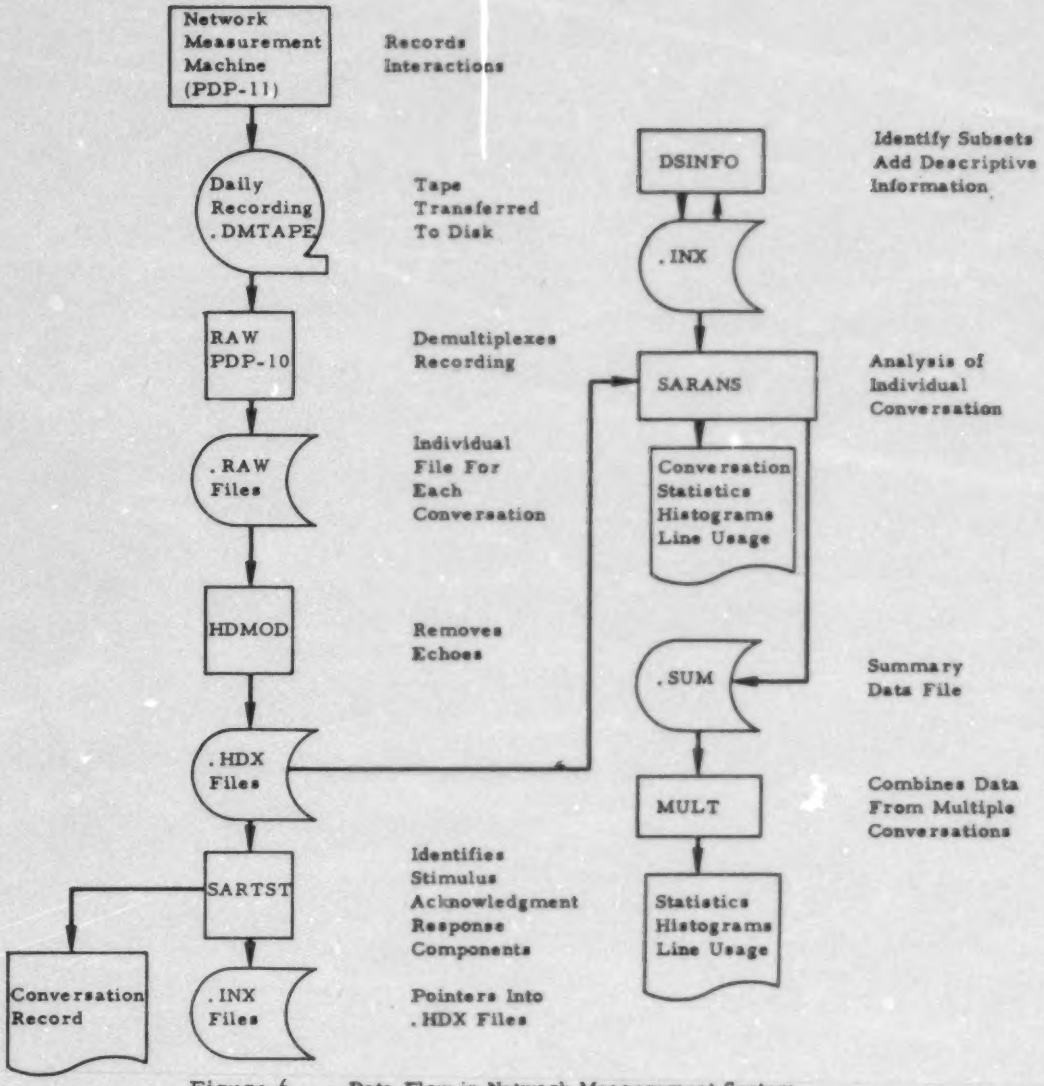


Figure 6. Data Flow in Network Measurement System

to determine conversation boundaries by semantic analysis. Where the system being used is known in advance, such analysis is generally successful. When the analysis fails, boundaries can be recognized by a human analyst and specified to the analysis system, or the data from several conversations can simply be left as one long conversation. Unless average conversation length is under investigation, this should not bias the results.

The demultiplexed conversations are referred to as "RAW" data files and are employed as input to further analysis. A separate RAW file is produced for each conversation. File labeling conventions reflect the recording session from which each file was produced and, where known, the system whose response was recorded.

At this point, the analysis programs begin to reflect the conceptual models. If requested, the RAW file may be run through a transformation to remove echos. Upper case echoes of lower case input can also be removed if requested. The data files produced by this process are referred to as "half-duplex" ("HDX") files.

In order to apply the SAR model to the collected data, the HDX file is then scanned to determine the boundaries of message groups and the stimulus, acknowledgement and response components. This information is recorded in an index file ("INX") for future reference. As part of the scanning, a conversational record is prepared, with the message groups and their constituent parts identified. In this record, the non-printing characters are optionally represented by mnemonics. For example, a carriage return is indicated by <CR> and blank spaces by <SP>. This permits an analyst to review the record of the conversation and easily determine exactly what characters were sent, either by the user or the system.

Optionally, the derived data parameters are printed for each message group. Various statistics are presented describing the entire conversation. For each derived data parameter, a frequency of occurrence histogram can be printed. The mean, standard deviation, 50th percentile (median), 90th percentile and 95th percentile are all calculated and printed. Additional data of possible interest which we present are the percentage of printing and non-printing characters for the user and the system, the effective character transmission rates for the user and the system, and the number of echo characters detected and deleted. Figure 7 illustrates some typical outputs from the analysis routines.

STIMULUS-RESPONSE DELAY TIME (IN SECONDS)

FREQUENCY CLASS BASE VALUE (NUMBER OF OCCURRENCES)
GROUP FREQUENCY RANGE 0.4
TOTAL NUMBER OF OCCURRENCES= 2358 (IN RANGE)
TOTAL NUMBER OF OCCURRENCES= 2402 (ABSOLUTE)
0 OCCURRENCE(S) BELOW THE MINIMUM, 44 OCCURRENCE(S) ABOVE THE MAXIMUM

MEAN=	1.2	STANDARD DEVIATION=		2.1	
SKEW=	4.3	KURTOSIS=		24.0	
MEDIAN=	0.7	90%=	2.6	95%=	4.8

Figure 7. Sample Output from Network Measurement System

LINE UTILIZATION STATISTICS VERSION 750523

300.0 BITS PER SECOND TRANSMISSION LINE RATE

	NUMBER	% OF TOTAL	% OF SUBTOTAL
TOTAL NUMBER OF CHARACTERS	307415	100	
USER	36838	11	100
PRINTING	32623	10	88
NON-PRINTING	4215	1	12
SYSTEM	270577	89	100
PRINTING	243070	10	89
NON-PRINTING	27507	10	11
SUBTOTAL # OF PRINTING CHARS	275693	89	100
USER	32623	10	11
SYSTEM	243070	79	89
SUBTOTAL # OF NON-PRTING CHRS	31722	11	100
USER	4215	1	13
SYSTEM	27507	10	87

LINE UTILIZATION (MEAN)
DURING TRANSMISSION ONLY (BURSTINESS)

USER	7.3%
ACKNOWLEDGEMENT	17.9%
RESPONSE	59.7%

LINE UTILIZATION (MEAN)

USER	3.9%
SYSTEM	46.7%

Figure 7. (continued) Sample Output
from Network Measurement System

Subsets

Frequently it may be desired to examine and compare different subsets of one or more conversations. Subsets of interest may be of many different types, and may or may not cross conversation boundaries. Temporal subsets, for example, might be used to investigate fatigue factors in long conversations. Subsets based on software system use could be used to compare the responsiveness of the various processors or the preference of users for particular ones. To satisfy such curiosity, the NMS provides the means for a measurement analyst to group data into subsets and for these subsets to be analyzed separately.

Subsets are produced by the analyst utilizing the conversation record which was produced along with the index file. It is necessary for the analyst to be able to recognize the members of the various subsets; for logical subsets this might require some knowledge of the computer system being used. Working with an interactive program, the analyst creates entries in the INX file, which in turn contains descriptive information about the contents of the HDX file, including identification of the user, the computer, and the date of recording. Additionally, the INX file contains pointers to the stimulus, acknowledgement and response of each message group along with the identification of the subset. Employing this information, it is possible to retrieve message groups according to subsets and logical combinations of subsets, and to perform the desired analysis.

USES OF THE NETWORK MEASUREMENT SYSTEM

The measurement techniques which have been described may be used in the design, selection, procurement and improvement of computer networks and network services. In the design of a computer system or network, the ability to measure external performance will be useful in specifying service design goals and in determining how well these goals are met. Objective measurement will allow the direct comparison of alternate designs by implementers and users alike. Thus, service providers can implement systems with specific design goals for service, and users can specify, perhaps contractually, the level of service they expect. Both buyer and seller should benefit from having a straightforward way to determine if contract terms are being met. Finally, these measurement techniques can be applied to installed systems to determine the level of service being provided, identify where

improvements are needed, and to evaluate the effects of changes made to the system.

Procurement

One of the most difficult parts of a procurement action for interactive or network services is the specification of requirements. Prospective purchasers seem to have great difficulty predicting and expressing the demand that terminal users are expected to place on the system, and equal difficulty in deciding what the response characteristics of the service should be. The Network Measurement System can be of use in resolving these difficulties by first measuring the current service and then the proposed new services.

Assuming that there is an interactive or network service in use already, the Network Measurement System can determine what the user demands are on that system and what response it is providing. Specification of the anticipated workload is an essential part of any requests for proposal [FE72]. Knowledge of current demand provides a starting point for forecasting future demand by providing a profile of user characteristics. Knowledge of current responsiveness, together with current user evaluations, provides the basis for strengthening or relaxing such requirements in the future.

Application of the NMS to proposed systems could be part of any benchmarking evaluation to determine if responsiveness is as claimed in the bid. After selection and procurement, the NMS could continually be used to ensure compliance with contractual provisions. This would require that performance requirements be specified in the terms of the contract -- a practice that is not too common today.

Operation

Applied to an existing network or service, the Network Measurement System can provide a variety of useful and interesting information. Similar functions have been served by internal measurement techniques [BO74], but the NMS has the distinct advantage of measuring systems directly without perturbing them.

One application is to learn about users and their characteristics. This may be interesting information in its own right to psychologists and human factors engineers, but it is essential to network designers and implementors. The average transmission rate of users, for example, is a key

factor in designing efficient multiplexing and concentrating communications networks. Similarly, knowledge of the relative utilization of various facilities by users can provide guidance as to where optimization efforts should be directed.

Another application is to learn about the network. As we will describe in the following section, it is possible to use the Network Measurement System to determine the delay introduced by a data communications system. This is one type of information about the network which should be of interest to users. Another area of interest is the overhead of communications protocols. The NMS can aid in the analysis of this area by simply providing the capability to print out the complete traffic on the communications line, including control characters. Through investigations of this type it may be possible to identify potential areas for system improvement.

Applications

Several immediate applications of the Network Measurement System are planned and, in fact, are in various stages of completion. These applications have been selected to provide information about both users of network services and the services themselves, as well as to provide a testbed for operational capabilities of the NMS. These applications have not been designed as rigorous, parametric experiments with control groups and the like. Such experiments are expected to follow.

The Network Measurement System was first used in sort of a "bootstrap" mode to record the conversations of the programmers who were implementing the data analysis programs for it. This application provided data to verify the operation of the data acquisition system, especially the time tagging procedure, and to use to test and debug the analysis routines.

The next application was to connect the NMS to a port of a Government computer system providing interactive bibliographic information retrieval services. This was done at the request of the agency providing the service to help them evaluate the communications system which was delivering their product to customers. Since the computer system was located several miles away from the NMS, it was necessary to design and install a special remote interface which, in concert with a leased line and pair of modems, could acquire the necessary data without interfering with the conversation in progress. On the software side, it was necessary to

program a special algorithm to recognize the beginning and end of conversations based on the particular semantics of logon and logoff, since individual users frequently changed (on the same line) without hanging up, and since the interface hardware being used couldn't detect hangup conditions even if they did. (The hardware is being changed).

Simply providing gross aggregate statistics for response time over a five month period was of considerable interest to the agency management, since there were changes in both operating system and hardware during this period. The more specific question of the degradation to response time introduced by the network delivery system was investigated in a special experiment. A single user terminal was simultaneously connected to two input ports of the computer system, but by different means. One port was connected to the terminal through the communications service in question; the other was connected through the dial telephone network. Since the inputs to each port were identical and generated at exactly the same time, it was possible to isolate the extra delay (beyond that of pure transmission time) introduced by the (store and forward) communications service.

The Network Measurement System is equipped with both dial-in ports accessible through the public switched telephone network and an automatic calling unit. Cooperating users can dial the NMS instead of their regular service. After identifying itself, the NMS requests that the user identify the desired system. When the user responds with a telephone number or a name (which must have been pre-stored in the NMS directory with the telephone number), the NMS will seek to dial that system. If a successful connection is made, the user is informed and may proceed in the normal manner. From this point on the NMS is completely transparent to both user and system.

A planned effort is the measurement of computer services under controlled conditions. When the services are delivered interactively, it is necessary to employ another computer if the stimuli are to be completely controlled and repeatable. The Network Measurement System will be used to collect data which can be transformed into scripts which can be used by a computer-based "measurement driver". It is anticipated that as part of a procurement action, each vendor could be supplied with a scenario which included a set of instructions to be applied to the proposed computer service. The instructions might include using an editor to create a source language program, iteratively compiling it and correcting errors, and executing the final object program with certain data. Vendor personnel would execute the instructions while

connected to the NMS. The recorded conversations could then be used by a measurement driver to exercise the computer service under various controlled conditions of loading. The NMS would again be used to record the conversations at this time, so that comparisons could be made between competing systems.

CONCLUSIONS

The quantitative analysis of interactive systems is performed at a very crude level of sophistication today. This is partially because it has a basis in the evaluation of computer systems in general, which is still a very inexact science, and partially because it deals with a system involving humans, whose performance is always difficult to assess.

A major problem in investigating these matters has been the lack of measurement instruments which are convenient to use and which do not perturb the system under investigation. We have constructed such a measurement instrument. The Network Measurement System is transparent to users and does not consume any network resources.

We have only just begun to use the Network Measurement System in a controlled and rigorous way. We have discussed some preliminary applications of the system, and some possible uses. We plan to apply the system in many of these areas as part of our general inquiry into the design, selection and use of computer networks.

FOOTNOTES

1. Although the following terms have defined meanings in communications technology, we employ them here in the broader context of computer networking conventions:

"Half-duplex" refers to the case where the user's keyboard is connected to his printer, and to the convention that when the computer is producing output it does not respond to input from the user (except perhaps to a special "attention" signal). When the computer is not producing output, it is in a mode to accept input from the user.

"Echo-plex" refers to the case where the user's keyboard is locally disconnected from the printer mechanism. Characters input by the user are returned

to his terminal s printer by the network. Echo-plex operation allows computer output to be interspersed with echoed user input.

"Full-duplex" also refers to the case where the user's keyboard and printer are disconnected, but it additionally implies that the echoed character need not be identical to the character sent (e.g., ETX ("control-C") may be echoed as the two character sequence "^C"). Furthermore, the echoing may be delayed in time from the receipt of the user's input The terms echo-plex and full-duplex are often used interchangeably when the distinction is unimportant, unknown, or misunderstood.

2. American Standard Code for Information Interchange, X3.4-1968.

BIBLIOGRAPHY

AB73 Abrams, M. D., Lindamood, G. E., and Pyke, T. N., Jr., "Measuring and Modeling Man-Computer Interactions," Proceedings of the 1st Annual SIGME Symposium on Measurement and Evaluation, February 1973, Pp. 136-142.

BE73 Bell, T. E., Boehm, B. W., and Watson, R. A., "How to Get Started on Performance Improvement," Computer Decisions, March 1973, Pp. 30-34.

BL74 Blanc, R. P., Review of Computer Networking Technology , NBS Technical Note 804, January 1974.

BOE71 Boehm, B. W., Seven, M. J., and Watson, R. A., "Interactive Problem Solving -- An Experimental Study of Lockout Effects," Proc. Spring Joint Computer Conference, 1971, Pp. 205-210.

BOI74 Boies, S. J., User Behavior on an Interactive Computer System, IBM System Journal, No. 1, 1974, Pp. 2-18.

CA68 Carbonell, J. R., Elkind, J. I., and Nickerson, R. S., "On the Psychological Importance of Time in a Time Sharing System," Human Factors, Vol. 10, No. 2, 1968, Pp. 135-142.

CO74 Cotton, Ira W. "Cost-Benefit Analysis of Interactive Systems," Proceedings of 2nd Jerusalem Conference on Information Technology, August 1974, pp. 729-746.

FE72 Ferrari, D., "Workload Characterization and Selection in Computer Performance Measurement," Computer, July/August 1972, Pp. 18-24.

FU70 Fuchs, E. and Jackson, P. E., "Estimates of Distributions of Random Variables for Certain Computer Communications Traffic Models," Communications of the ACM, Vol. 13, No. 12, December 1970, Pp. 752-757.

GR75 Grubb, Dana S. and Cotton, Ira W., Requirements of Information Processing Systems for Supporting Telecommunications Services, NBS Technical Note, in press.

JA69 Jackson, P.E. and Stubbs, C. D., "A Study of Multiaccess Computer Communications," Proc. Spring Joint Computer Conference, 1969, Pp. 491-504.

JU71 Jutila, S. T. and Baram, G., "A User-Oriented Evaluation of a Time-Shared Computer System," IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-1, No. 4, October 1971, Pp. 344-349.

KA69 Kamerman, A., Establishing and Measuring Conversational Performance Objectives for Time Sharing Systems, Technical Report TR53.015, IBM Systems Development Division, Yorktown Heights, N.Y., June 1969.

MIe72 Miller, E.F., Jr., "Bibliography on Techniques of Computer Performance Analysis," Computer, September/October 1972, Pp. 35-47.

MIr68 Miller, R. B., "Response Time in Man-Computer Conversational Transactions," Proc. Fall Joint Computer Conference, 1968, Pp. 267-277.

PR73 Proceedings of the 1st Annual SIGME Symposium on Measurement and Evaluation, ACM Order Department, February 1973.

RO75 Rosenthal, R., Rippy, D. E. and Wood, H., The Network Measurement Machine -- A Data Collection Device for Measuring the Performance and Utilization of Computer Networks, NBS Technical Note, in press.

ST72 Streeter, D.N., "Cost-Benefit Evaluation of Scientific Computing Services," IBM Systems Journal, No. 3, 1972, Pp. 219-233.

WA75 Watkins, S. W. and Abrams, M. D., Data Treatment and Analysis in the Measurement of Computer Networks, NBS Technical Note, in press.

NBS-114A (REV. 7-73)

U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET	1. PUBLICATION OR REPORT NO. NBS TN-880	2. Gov't Accession No.	3. Recipient's Accession No.

4. TITLE AND SUBTITLE

The Service Concept Applied to Computer Networks

5. Publication Date August 1975

6. Performing Organization Code

7. AUTHOR(S) Marshall D. Abrams, Ira W. Cotton

8. Performing Organ. Report No.

9. PERFORMING ORGANIZATION NAME AND ADDRESS

NATIONAL BUREAU OF STANDARDS
DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20234

10. Project/Task/Work Unit No. 6502120

11. Contract/Grant No.

12. Sponsoring Organization Name and Complete Address (Street, City, State, ZIP)

13. Type of Report & Period Covered Interim

14. Sponsoring Agency Code

15. SUPPLEMENTARY NOTES

Library of Congress Catalog Card Number: 75-600056

16. ABSTRACT *(A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)*

The Network Measurement System (NMS) represents the implementation of a new approach to the performance measurement and evaluation of computer network systems and services. By focusing on the service delivered to network customers at their terminals, rather than on the internal mechanics of network operation, measurements can be obtained which are directly relevant to user needs and management concerns. Furthermore, the type of measurement necessary to implement this approach can be made directly, without perturbing the network system under test.

This technical note introduces the service concept and other background information necessary to understand the need for and use of the NMS. The fundamental distinction between service and internal efficiency is clarified, both in general and in the environment of computer networks. A number of different measures of service are then discussed, followed by the presentation of several models of interactive use of networks. Then, the practical aspects of gathering data and applying this information to service measurement are reviewed, leading to a presentation of the NMS as it is implemented. The note ends with a discussion of applications for the NMS.

17. KEY WORDS *(six to twelve entries; alphabetical order; capitalize only the first letter of the first key word unless a proper name; separated by semicolons)*

Computer networks; interactive computing; man-machine interaction; performance measurement; time-sharing

18. AVAILABILITY	19. SECURITY CLASS (THIS REPORT)	21. NO. OF PAGES
[X] Unlimited	UNCLASSIFIED	38
[ ] For Official Distribution. Do Not Release to NTIS		
[X] Order From Sup. of Doc., U.S. Government Printing Office Washington, D.C. 20402, SD Cat. No. C13. 46:880	20. SECURITY CLASS (THIS PAGE)	22. Price
[ ] Order From National Technical Information Service (NTIS) Springfield, Virginia 22151	UNCLASSIFIED	85 cents

OMM-DC 29042-P74

NBS TECHNICAL PUBLICATIONS

PERIODICALS

JOURNAL OF RESEARCH reports National Bureau of Standards research and development in physics, mathematics, and chemistry. It is published in two sections, available separately:

• Physics and Chemistry (Section A)

Papers of interest primarily to scientists working in these fields. This section covers a broad range of physical and chemical research, with major emphasis on standards of physical measurement, fundamental constants, and properties of matter. Issued six times a year. Annual subscription: Domestic, $17.00; Foreign, $21.25.

• Mathematical Sciences (Section B)

Studies and compilations designed mainly for the mathematician and theoretical physicist. Topics in mathematical statistics, theory of experiment design, numerical analysis, theoretical physics and chemistry, logical design and programming of computers and computer systems. Short numerical tables. Issued quarterly. Annual subscription: Domestic, $9.00; Foreign, $11.25.

DIMENSIONS/NBS (formerly Technical News Bulletin)—This monthly magazine is published to inform scientists, engineers, businessmen, industry, teachers, students, and consumers of the latest advances in science and technology, with primary emphasis on the work at NBS. The magazine highlights and reviews such issues as energy research, fire protection, building technology, metric conversion, pollution abatement, health and safety, and consumer product performance. In addition, it reports the results of Bureau programs in measurement standards and techniques, properties of matter and materials, engineering standards and services, instrumentation, and automatic data processing.

Annual subscription: Domestic, $9.45; Foreign, $11.85.

NONPERIODICALS

Monographs—Major contributions to the technical literature on various subjects related to the Bureau's scientific and technical activities.

Handbooks—Recommended codes of engineering and industrial practice (including safety codes) developed in cooperation with interested industries, professional organizations, and regulatory bodies.

Special Publications—Include proceedings of conferences sponsored by NBS, NBS annual reports, and other special publications appropriate to this grouping such as wall charts, pocket cards, and bibliographies.

Applied Mathematics Series—Mathematical tables, manuals, and studies of special interest to physicists, engineers, chemists, biologists, mathematicians, computer programmers, and others engaged in scientific and technical work.

National Standard Reference Data Series—Provides quantitative data on the physical and chemical properties of materials, compiled from the world's literature and critically evaluated. Developed under a world-wide program coordinated by NBS. Program under authority of National Standard Data Act (Public Law 90-396).

NOTE: At present the principal publication outlet for these data is the Journal of Physical and Chemical Reference Data (JPCRD) published quarterly for NBS by the American Chemical Society (ACS) and the American Institute of Physics (AIP). Subscriptions, reprints, and supplements available from ACS, 1155 Sixteenth St. N. W., Wash. D. C. 20056.

Building Science Series—Disseminates technical information developed at the Bureau on building materials, components, systems, and whole structures. The series presents research results, test methods, and performance criteria related to the structural and environmental functions and the durability and safety characteristics of building elements and systems.

Technical Notes—Studies or reports which are complete in themselves but restrictive in their treatment of a subject. Analogous to monographs but not so comprehensive in scope or definitive in treatment of the subject area. Often serve as a vehicle for final reports of work performed at NBS under the sponsorship of other government agencies.

Voluntary Product Standards—Developed under procedures published by the Department of Commerce in Part 10, Title 15, of the Code of Federal Regulations. The purpose of the standards is to establish nationally recognized requirements for products, and to provide all concerned interests with a basis for common understanding of the characteristics of the products. NBS administers this program as a supplement to the activities of the private sector standardizing organizations.

Federal Information Processing Standards Publications (FIPS PUBS)—Publications in this series collectively constitute the Federal Information Processing Standards Register. Register serves as the official source of information in the Federal Government regarding standards issued by NBS pursuant to the Federal Property and Administrative Services Act of 1949 as amended, Public Law 89-306 (79 Stat. 1127), and as implemented by Executive Order 11717 (38 FR 12315, dated May 11, 1973) and Part 6 of Title 15 CFR (Code of Federal Regulations).

Consumer Information Series—Practical information, based on NBS research and experience, covering areas of interest to the consumer. Easily understandable language and illustrations provide useful background knowledge for shopping in today's technological marketplace.

NBS Interagency Reports (NBSIR)—A special series of interim or final reports on work performed by NBS for outside sponsors (both government and non-government). In general, initial distribution is handled by the sponsor; public distribution is by the National Technical Information Service (Springfield, Va. 22161) in paper copy or microfiche form.

Order NBS publications (except NBSIR's and Bibliographic Subscription Services) from: Superintendent of Documents, Government Printing Office, Washington, D.C. 20402.

BIBLIOGRAPHIC SUBSCRIPTION SERVICES

The following current-awareness and literature-survey bibliographies are issued periodically by the Bureau:

Cryogenic Data Center Current Awareness Service

A literature survey issued biweekly. Annual subscription: Domestic, $20.00; foreign, $25.00.

Liquefied Natural Gas. A literature survey issued quarterly. Annual subscription: $20.00.

Superconducting Devices and Materials. A literature survey issued quarterly. Annual subscription: $20.00. Send subscription orders and remittances for the preceding bibliographic services to National Technical Information Service, Springfield, Va. 22161.

Electromagnetic Metrology Current Awareness Service Issued monthly. Annual subscription: $100.00 (Special rates for multi-subscriptions). Send subscription order and remittance to Electromagnetics Division, National Bureau of Standards, Boulder, Colo. 80302.